A genetic epidemiology approach to cyber-security

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPUTER SCIENCE
SOFTWARE

Received
10 April 2014

Accepted
17 June 2014

Published
16 July 2014

Correspondence and requests for materials should be addressed to S.G. (sg.ccnr@gmail.com)

A genetic epidemiology approach to cyber-security

Santiago Gil[1,2], Alexander Kott[3] & Albert-László Barabási[1,4,5]

[1]Center for Complex Network Research, Northeastern University, Boston, MA 02130, USA, [2]Seed Scientific, New York, NY 10013, [3]Network Science Division, Army Research Laboratory, Adelphi, MD 20783, USA, [4]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA, [5]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

While much attention has been paid to the vulnerability of computer networks to node and link failure, there is limited systematic understanding of the factors that determine the likelihood that a node (computer) is compromised. We therefore collect threat log data in a university network to study the patterns of threat activity for individual hosts. We relate this information to the properties of each host as observed through network-wide scans, establishing associations between the network services a host is running and the kinds of threats to which it is susceptible. We propose a methodology to associate services to threats inspired by the tools used in genetics to identify statistical associations between mutations and diseases. The proposed approach allows us to determine probabilities of infection directly from observation, offering an automated high-throughput strategy to develop comprehensive metrics for cyber-security.

Given the extent to which critical infrastructures and day-to-day human and economic activity are dependent on secure information and communication networks, we must understand their weaknesses and uncover the risks that threaten their integrity[1–3]. Much work has been devoted in the past years to understanding the global risks resulting from node removal in communication and infrastructure networks[3–6], or posed by viruses and malware[7]. Less is known, however, about the specific risks that pertain to compromising single nodes.

As cyber-threats proliferate both in their volume and complexity, new strategies are necessary to prevent intrusions and to cope with their impact[8]. Current research in cyber-security focuses on characterizing and modelling specific attacks, aiming to understand the mechanisms of infiltration, detection and mitigation[9–17]. Despite significant efforts in this direction, the number of attacks continues to increase each year[18,19]. The constantly increasing vulnerabilities and intrusions have lead to a paradigm shift in security: it is now widely recognized that absolute security is unattainable[20,21]. While very effective methods to detect and aggregate software vulnerabilities have been developed, determining actual risks and probabilities of compromise largely relies on expert opinion and heuristics[22–25]. Instead, we need systematic methods to estimate the magnitude of the potential risk by determining the probability that a system is compromised and the risk associated with specific settings and configurations[26]. Understanding the nature and the components of the current risks could help us devote resources to most efficiently reduce the chance of intrusion.

In this work we introduce an unbiased, context-free and high-throughput statistical framework to evaluate the susceptibility of a computer to individual threats. By combining infection data from threat logs with network scans we identify correlations between the network services detected on a host and the threats affecting it. The proposed method offers a systematic framework to identify new attack vectors and to evaluate the role of individual factors contributing to the susceptibility to each threat.

Data collection

To understand the recurring threats experienced by a large computer network, we collected the threat logs of an Intrusion Detection and Prevention System (IDS/IPS) protecting a university network of approximately 30, 000 hosts. The IDS/IPS monitors incoming and outgoing traffic, searching for matching signatures from a threat database. Each positive match is logged with a time stamp, source and target IP addresses as well as the name and identification number of the detected threat. A small fraction of the detected threats (~3.2%) originate *within* the network. Since the threats in the database correspond to well documented malicious activity and constitute violations of security policies established by the network administrators, we assume that a threat originating within the University network is evidence that the corresponding host at the source IP address has been compromised. Our

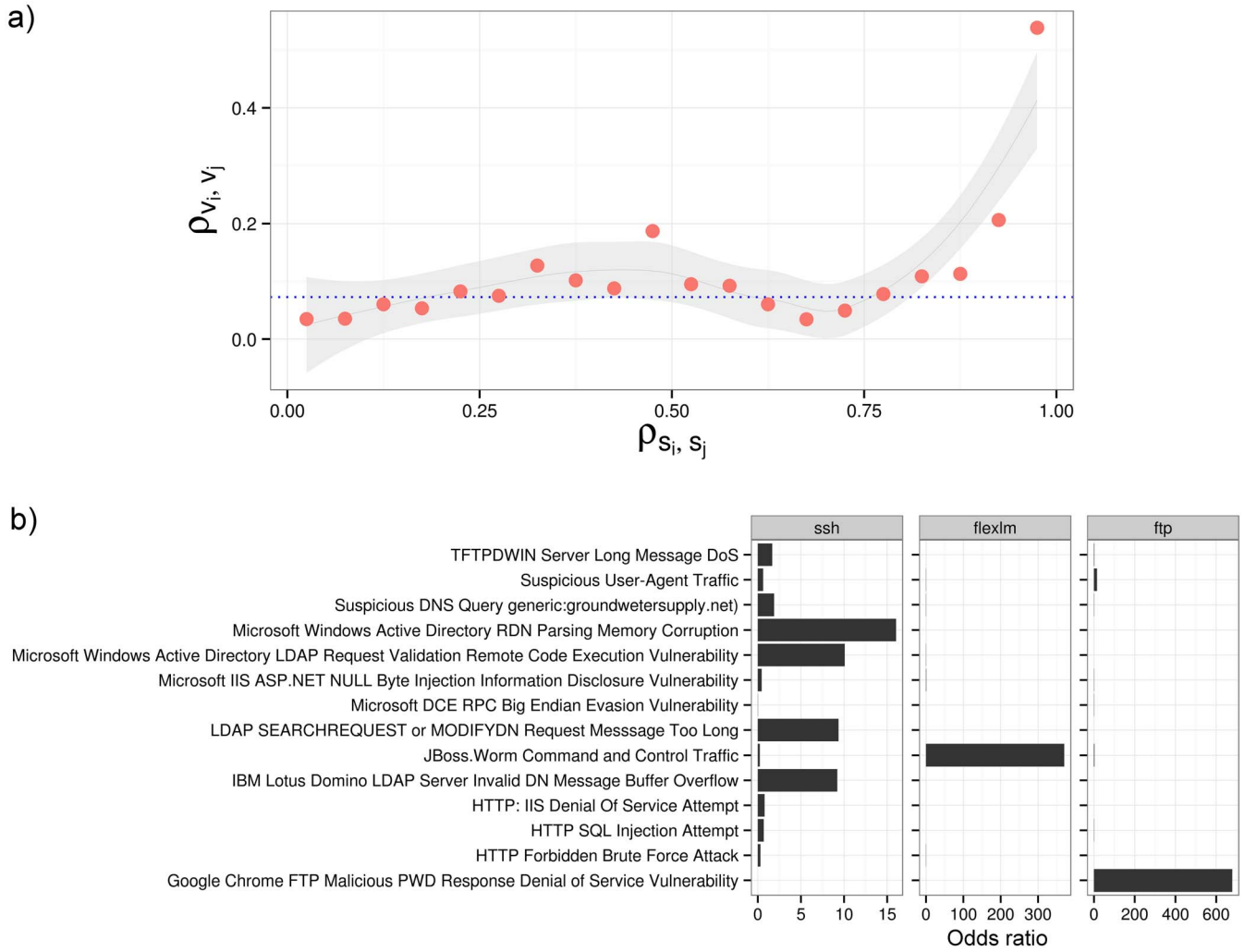

Figure 1 | Correlating threats and services. (a) Pairs of hosts running similar services have an increased likelihood of being compromised by the same threats. The plot shows the mean value of the correlation between threat profiles as a function of the correlations between services for all pairs of hosts, indicating that the likelihood is maximal for computers running the same services. The blue line represents the expectation value for random resampling of the data. As the Pearson correlation is not defined for a constant signal, the plot excludes hosts for which no threats or no services have been found. (b) Odds ratios for the most common threats in the presence of three services: *ssh*, *flexlm* and *ftp*. The presence of certain services drastically increases the odds of being compromised by specific threats.

threat-dataset consists of 501 days of logs produced by the IDS/IPS beginning in September 2012, with approximately 16 million entries.

To remotely gather information about hosts in the network, we also performed full network scans of all IP addresses within the university, obtaining the list of network services running on all hosts found. A total of 11, 726 hosts were successfully scanned, identifying 464 distinct services. In this sample, 10, 560 hosts were found in the logs to have been compromised by at least one of 278 distinct threats.

Results

Correlations between services and threats. Our starting hypothesis is that the probability that a host is compromised by a specific threat is largely determined by the set of reachable services running on it[27]. Indeed, many attacks take advantage of vulnerabilities in specific network services to bypass authentication and gain access[28]. Additionally, certain types of malware may autonomously scan for services to identify potential targets, or even actively open ports after infection. Furthermore, the presence of specific services on a host can be indicative of the ways in which the host is used, its intended purpose or its user's security awareness, allowing us to estimate the host's susceptibility to specific threats.

Evidence for this hypothesis is provided in Figure 1a, showing the correlations between services and threats. We define the threat profile of host i as vector $\mathbf{v}^i$, where v^i_k represents the fraction of log entries for host i that correspond to threat k, and its service conformation as $\mathbf{s}^i$, where s^i_l is the number of instances of service l running on host i. We calculate $\rho(\mathbf{v}^i, \mathbf{v}^j)$ in function of $\rho(\mathbf{s}^i, \mathbf{s}^j)$ for all pairs of hosts i and j, where ρ is the Pearson correlation. The blue line is the expectation value obtained under randomization of the data that removes all potential correlations between services and threats. We find that, on average, the correlation between threat profiles for most host pairs remains around the values expected in the random case. However, for host pairs with high correlation between their services ($\rho(\mathbf{s}^i, \mathbf{s}^j) > 0.8$) the average correlation between their threat profiles increases sharply. This means that hosts running similar services have a significantly increased likelihood of being compromised by the same threats. This likelihood is maximal for hosts with identical services ($\rho(\mathbf{s}^i, \mathbf{s}^j) = 1$).

Figure 1b shows three examples supporting the hypothesis that the presence of specific network services is a determining factor for the susceptibility to certain threats. Indeed, we find that a computer running the *ssh* service is 16 times more likely to be flagged for the

Active Directory RDN Parsing Memory corruption than computers without that service. More dramatically, a host running the *ftp* service is 678 times more likely to trigger alerts for Chrome's FTP PWD response vulnerability, and hosts running *flexlm* are about 370 times more likely to be infected by the JBoss Worm (CVE-2010-0738 in the Common Vulnerabilities and Exposures Database, www.cve.mitre.org).

Case-control study. The ability of a threat to compromise a host depends on a combination of factors, from the software running on the host and its level of exposure to the user's behavior and security awareness, to name only a few. Yet, the finding that hosts running similar services are susceptible to similar threats suggests that significant aspects of these factors are represented in the specific combination of network services present on a host. Therefore, to understand a system's risks we must systematically untangle the correlations between individual services and threats. This is traditionally done through domain experience and contextual knowledge—system administrators must constantly stay informed about current trends pertaining to their system and the type of services they run. However, given the rapidly evolving complexity of both threats and services, this is a difficult and inefficient task. Next we show that we can automate this process by relying on tools borrowed from genetics.

We structure our analysis as a standard case-control study. We consider a binary state model in which a given threat has either been detected on a host or not, and the host is either running a given network service or not. Our case-hosts for the study of a specific threat are therefore all the hosts on which that threat has been detected by the IDS/IPS. All hosts on which the threat has not been observed serve as the control group.

For a given threat k, the group of case-hosts A_k that have been infected by the threat and the group of control hosts C_k are given by

$$A^k = \{i : v_k^i > 0\} \tag{1}$$

$$C^k = \{i : v_k^i = 0\}. \tag{2}$$

For each service these two groups are split into hosts that are running the service and hosts that are not. Since we consider each service as a risk factor, we refer to hosts running a specific service as *exposed*. For a given service l under scrutiny, the different groups $A_k = {}^k_{l+} \cup A^k_{l-}$ and $C_k = C^k_{l_+} \cup C^k_{l-}$ are given by

$$A^k_{l_+} = \{i \in A^k : s_l^i > 0\} \tag{3}$$

$$A^k_{l-} = \{i \in A^k : s_l^i = 0\} \tag{4}$$

for the exposed and unexposed case-hosts, respectively, and

$$C^k_{l_+} = \{i \in C^k : s_l^i > 0\} \tag{5}$$

$$C^k_{l-} = \{i \in C^k : s_l^i = 0\} \tag{6}$$

for the exposed and unexposed controls, where the symbols + and − correspond to the service running and not running on the hosts, respectively. With this, we construct a 2 × 2 contingency table as shown in Table I.

Determining whether service l is significantly associated with threat k requires us to evaluate the statistical differences between the two columns of this table. This association problem is equivalent to that encountered in genetic epidemiology, where the goal is to assess the impact of specific genetic mutations on complex diseases[29], a problem addressed using genome-wide association studies (GWAS)[30]. In these studies, a large number of genetic markers along the entire genome are compared between a population of healthy and disease-affected individuals. GWAS offers a high-throughput method for the unbiased identification of mutation-genotype associations.

Table I | Case-control framework. Contingency table for the evaluation of an association between the presence of a given network service and the infection by a given threat

	running the service	not running the service
infected by the threat	$\lvert A^k_{l+}\rvert$	$\lvert A^k_{l-}\rvert$
threat not detected in host	$\lvert C^k_{l+}\rvert$	$\lvert C^k_{l-}\rvert$

We equate the set of all scanned network services to the set of chosen mutations (biomarkers) in a GWAS. The presence of a given service on a host corresponds to the presence of a certain allele (genetic variant) for a given gene. Therefore, our problem has an identical mathematical formulation to the problem of genetic association, in which case patients play the role of hosts that have been compromised by a specific threat of interest. A statistical association is then evaluated by calculating the statistical significance (p-value) of the discrepancy between the distribution of variants for each gene in the population of individuals with the disease and the distribution of the same variants in a population of healthy individuals[31].

In our model, we measure the risk posed by a service, regardless of how many instances are running on a given host. In genetics, this corresponds to a common dominant model of genetic penetrance, such that the presence of a given (dominant) variant of a gene is sufficient to cause an associated phenotype, regardless of the number of copies present. We resort to the Fisher exact test to evaluate the statistical significance of the associations between threats and services. In the case of GWAS, logistic regressions are also commonly used to evaluate statistical significance when the considered genetic model allows for different values of the penetrance (e.g. additive or multiplicative) for the different possibilities of homozygous and heterozygous cases. The Fisher test used here assumes that the distributions of exposed and unexposed hosts are statistically identical for both the case and the control groups. In other words, all computers are equally likely to be compromised by the threat, regardless of whether they are running or not the considered service. Therefore, any discrepancies in the proportions of compromised hosts between exposed and non-exposed hosts should be due to chance. The test provides the probability of observing the proportions measured in the data under the null hypothesis. The outcome is the p-value for the significance of the association, which, if smaller than an established threshold, indicates that the null hypothesis is false. Therefore, in this case there is a statistically significant difference in the likelihood of falling victim to threat k between hosts that are running service l and hosts that are not.

This allows us to construct a standard Manhattan plot for each threat, illustrated in Fig. 2 for three different threats. The plots show the negative logarithm of the p-value for each service with the threshold for significance set at 0.01, multiplied by an additional Bonferroni correction. Each dot is colored according to the sign of the correlation: red dots above the threshold indicate a positive association between the service and the threat; green dots denote negative associations. The resulting plot summarizes the risk profile of a threat.

Positive associations imply that the service in question constitutes a risk factor for computers running the specific service. This could be the result of one of two possible scenarios: a) direct association, in which the service is directly responsible for the host being compromised, b) indirect association, where the presence of the associated service correlates with the factors that are the true cause of the host's infection (note that associations can also be false-positives due to

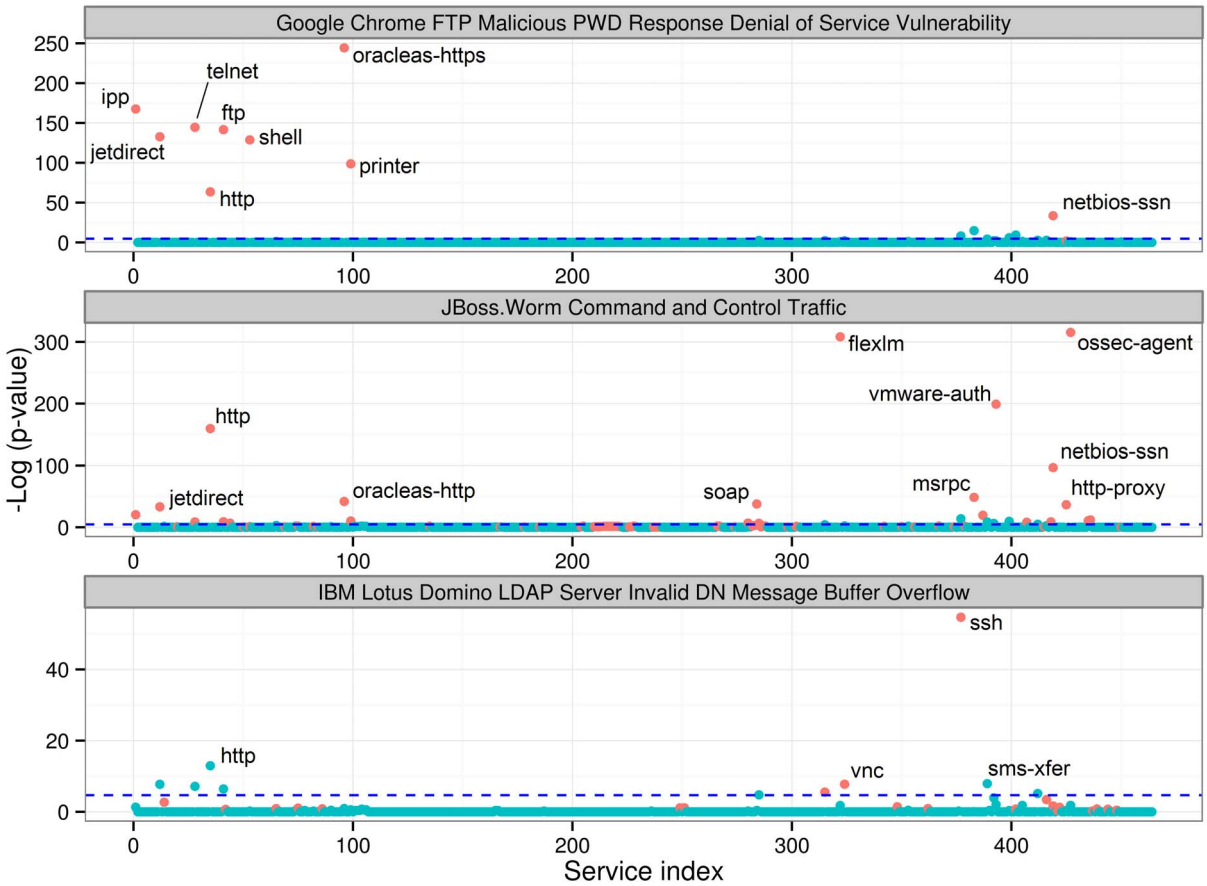

Figure 2 | Using GWAS to identify threat-service associations. Manhattan plots for three threats. For each service along the horizontal axis, the negative logarithm of the association p-value is displayed vertically. The dotted line denotes the threshold for statistical significance. The color of each dot corresponds to the sign of the logarithm of the corresponding odds ratio. Red dots above the threshold represent the services that are positively associated with the corresponding threat. Green dots indicate that the service's absence correlates with the threat. The ordering of the services along the horizontal axis is aleatory.

systematic biases in the data. We find no particular reason to suspect that this is at play in our data). Negative associations indicate that the service is significantly unlikely to be present on compromised hosts. In such cases, it is the *absence* of the service which correlates with the cause of the intrusion. Hence, the specific service may play an effective "protective" role.

Since network services represent a small subset of a system's characteristics, some of the identified associations will be indirect. An example is given by the JBoss Worm in Figure 2b. This threat affects the Red Hat Java Application Server, which typically runs behind a web server. Accordingly, the service *http* is present in an average of 2.55 instances in compromised hosts, while only 0.4 times in those unaffected. *vmware-auth* and *flexlm*, which are respectively 23 and 164 times more likely to be present in infected computers, also indicate that these hosts are part of a virtualized network that makes use of the application server. Even though the presence of the *ossec-agent* (a host-based intrusion detection system 354 times more likely to be found in compromised hosts) is intended to mitigate intrusions, it is also an indication that these are servers maintained by administrators, consistent with the presence of a JBoss Application Server. The association of all these services with the JBoss worm at high significance constitutes an indirect association through their co-occurrence with the actual causal factor.

Nevertheless, given that network services are the point of entry for many viruses and network intrusions, we expect many of these associations to be direct. For example, the threat ASP.NET NULL Byte Injection Information Disclosure Vulnerability (CVE-2007-0042, CVE-2011-3416) shows high significance for the association with the *remoting* service, which mediates communications in the .NET Framework. Since this service is used by ASP.NET for communications between server and client, it could plausibly provide an entry point for the potential attack, and thus constitute a direct association. Since there are alternative setups for this system (for example, using Windows Communication Foundation) blocking traffic for this service could have a measurable impact on the network's susceptibility to this threat, which would provide experimental evidence for this association.

Validation

The group of services whose p-value for a given threat lay above the statistical threshold are risk factors, implying that hosts running these services have a significantly higher likelihood of being compromised by the threat. To assess the predictive power of the proposed procedure, we divide our hosts into two cohorts of equal size, chosen randomly. We use one cohort as a calibration sample, identifying the risk factors (services) associated with each threat. The other cohort is used to measure the effect of running the associated risk-factor services on the likelihood of being compromised by each threat. Specifically, after obtaining the list of associated risk-factor services from the calibration cohort, we identify the hosts in the test cohort which are running at least one of the associated services (exposed hosts). We then compare the proportion of hosts that have been compromised by the threat amongst the exposed hosts with the proportion amongst unexposed hosts. This allows us to establish the probability that the observed proportions stem from the same

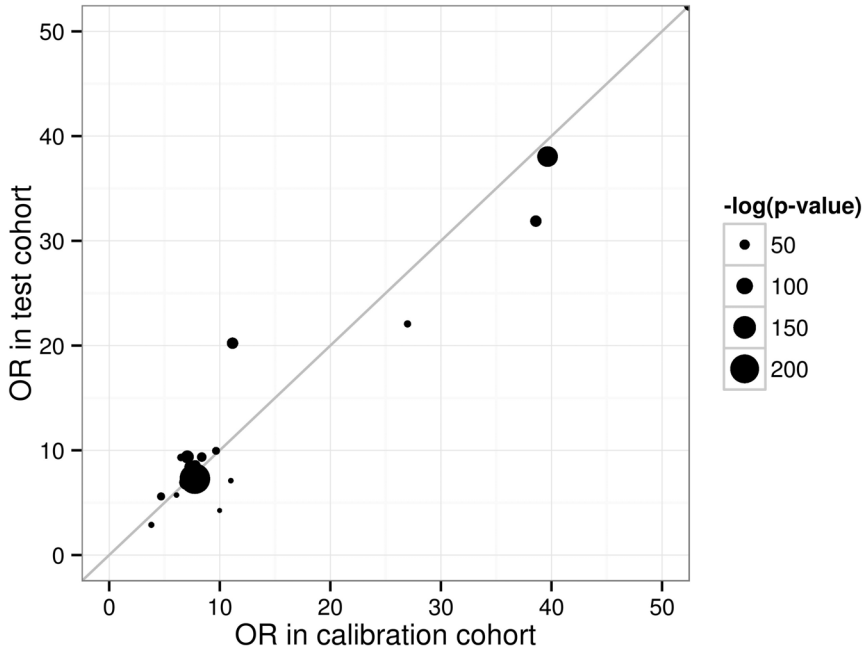

Figure 3 | Validating the GWAS predictions. Hosts running risk factor services are predicted to have significantly increased likelihoods of infection. These predictions are also valid for hosts that were excluded from the calculations. The plot shows the odds ratios corresponding to the predicted vulnerable hosts for each threat, measured in both the calibration and the test cohorts. The agreement between the two cohorts indicates that the information obtained from one sample can be used to make reliable predictions about other computers.

distribution. The resulting p-value corresponds to the probability that any distinction between exposed and unexposed hosts in the test cohorts are due to chance, or equivalently, the probability that the associated services obtained have no predictive power over the probability of infection. As an example, consider the case of the JBoss worm. Analyzing the calibration cohort, we find that the services most significantly associated with the worm are *vmware-auth, netbios-ssn, flexlm* and *ossec-agent*. In the test cohort, 641 hosts are running at least one of these three services, 119 of which (18.5%) were compromised by the threat. Amongst the remaining 5222 hosts that are not running any of these services, only 104 (1.9%) were compromised by the threat, a factor of ten difference. Thus, with a p-value of 2.44×10^{-58}, hosts running any of these three services have significantly increased likelihood of being infected by the JBoss worm, indicating the effectiveness of the method to detect susceptible hosts.

For 24 threats we successfully identified at least one risk-factor service, finding that 17 of these show statistical significance in the validation. The extent to which corresponding predictions about the incidence of threats in exposed hosts are reliable can be evaluated by comparing the odds ratios in both cohorts for each group of exposed hosts. In Figure 3 we show the odds ratio of each threat as measured in each cohort. In both cases, exposed hosts are those running at least one of the associated risk-factor services, obtained using data from the calibration cohort only. Although some variability between cohorts is observed, there is good agreement between the two cohorts, meaning that predictions made on the basis of one group of hosts can be carried over reliably to other hosts.

Table II | Host protection and threat reduction. If we can ensure the safety of a subset of the network by stringent control (or limit risk by blocking access to the hosts), we can calculate the resulting reduction in the incidence of a given threat. The desired degree of reduction in threat incidence determines a threshold for selecting hosts according to the risk factor services they run. This table shows the expected reductions in the incidence of the JBoss worm by means of different selection criteria

Number of risk-factors	Covered hosts	Threat reduction
at least 1	66.6%	98.9%
at least 5	14%	79.5%
at least 9	2.1%	45.1%
at least 12	1.7%	40%

Implementations

As we have shown above, the proposed statistical framework can be used to make predictions about the threat profile of a computer. If a host is automatically scanned when it connects to the network, the method allows us to immediately determine the list of threats by which the computer has an increased likelihood of being compromised. Therefore, packet inspection and traffic control can be tailored accordingly, offering a more efficient allocation of resources.

As for prophylactic measures at the network level, this method can be used to design firewall rules with reliable information about their potential effects on the probabilities of compromise. By identifying all the hosts running a high risk-factor service, one can concentrate resources on enacting stringent controls over a minimal number of hosts to obtain a maximal reduction in threat incidence. For the example of the JBoss worm, we identify the service *ossec-agent* as the most significantly associated risk-factor and having the highest odds ratio. This is a surprising finding given that the purpose of this software is to protect the host from intrusions, and it illustrates how this approach can help find unexpected correlations. Only 1.9% of the computers in the network run this service. Ensuring security against the JBoss Worm for these hosts alone could reduce the incidence of this threat by as much as 47%. Considering combinations of services allows us to fine-tune the degree of acceptable interventions to obtain a desired threat-reduction. For example, in Table II we show the percentage of computers one needs to effectively secure to attain different percentages of reduction in incidence of the JBoss worm if we consider hosts that are running a minimum num-

ber of associated risk-factor services. Adopting a selection criterion carries a trade-off between the scope of an intervention and the resulting threat reduction.

Of course, infallible prevention can hardly be achieved without resorting to drastic interventions, such as removing or otherwise isolating the host from the network. With a better understanding of the causality between threats and services, much less intrusive measures could be implemented with comparable results.

Discussion

In this paper we introduced a novel way to identify and quantify the susceptibility of individual computers to cyber-threats in terms of the network services that they run. We do so by establishing a direct mapping between the threat susceptibility problem and that of identifying statistical associations between genes and diseases in humans. The adopted approach relies on the methodologies developed in GWAS, aiming to associate genes with specific diseases.

Efficient and sophisticated tools to detect software vulnerabilities in a host are already available and used widely. However, they are limited to known and well-documented vulnerabilities, and their ability to evaluate the risks of compromise rely on heuristic metrics. The statistical approach developed here allows us to identify associations without preconceived assumptions about the underlying mechanisms. We can therefore explore the whole range of the variables that we consider (in this case, network services) in search for risk factors that may have been overlooked before. Such "agnostic", data-driven approaches are particularly valuable for making predictions in real-world scenarios, as they bypass the need for mechanistic explanations and contextual knowledge. Since we consider effective frequencies of infection *in vivo*, not only technical factors, but also behavioral, environmental and exogenous factors are automatically taken into account. For this reason, the proposed method offers reliable predictions about the effects of specific changes in the network in real-world environments.

The basis for this method resides in the observation that the network services running on a host play a defining role in its susceptibility to certain threats. This is not to deny that other variables may be equally important. Note, however, that our framework is not limited to services: if the appropriate data on other risk factors becomes available (*e.g.*, operating system, accounts and privileges, hardware and peripheral devices, applications, drivers, manufacturer, etc.), our method can determine their role for specific threats. Nevertheless, it is remarkable that such high statistical significance is observed using only services as risk factors. In contrast to other methods to assess risks that involve in-depth scanning, for which the collection of data can often be intrusive and even disruptive, the kind of data that our approach requires can be easily and systematically collected.

1. Choo, K.-K. R. The cyber threat landscape: Challenges and future research directions. *Com- puters & Security* **30**, 719–731 (2011).URL http://www.sciencedirect.com/science/ article/pii/S0167404811001040.
2. Brenner, J. *America the Vulnerable: Inside the New Threat Matrix of Digital Espionage, Crime, and Warfare* (Penguin. com, 2011).
3. Pastor-Satorras, R. & Vespignani, A. *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, 2007). URL http://books.google.com/books?id=EiySN0V4T_0C.
4. Cohen, R., Erez, K., ben Avraham, D. & Havlin, S. Breakdown of the internet under inten-tional attack. *Phys. Rev. Lett.* **86**, 3682–3685 (2001). URL http://link.aps.org/doi/10.1103/PhysRevLett.86.3682.
5. Albert, R., Jeong, H. & Barabasi, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000). URL http://dx.doi.org/10.1038/35019019.
6. Caldarelli, G. *Scale-Free Networks: Complex Webs in Nature and Technology*. No. 9780199211517 in OUP Catalogue (Oxford University Press, 2007).URL http://ideasrepec.org/b/oxp/obooks/9780199211517.html.
7. Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001). URL http://link.aps.org/doi/10.1103/PhysRevLett.86.3200.
8. Wulf, W. A. & Jones, A. K. Reflections on cybersecurity. *Science* **326**, 943–944 (2009).URL http://www.sciencemag.org/content/326/5955/943.short http://www.sciencemag.org/content/326/5955/943.full.pdf.
9. Ten, C.-W., Manimaran, G. & Liu, C.-C. Cybersecurity for critical infrastructures: Attack and defense modeling. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* **40**, 853–865 (2010).
10. Jajodia, S., Noel, S. & O'Berry, B. Topological analysis of network attack vulnerability. In Ku- mar, V., Srivastava, J. & Lazarevic, A. (eds.) *Managing Cyber Threats*, vol. 5 of *Massive Com- puting*, 247–266 (Springer US, 2005). URL http://dx.doi.org/10.1007/0-387-24230-9_9.
11. Wang, L., Islam, T., Long, T., Singhal, A. & Jajodia, S. An attack graph-based probabilistic security metric. In Atluri, V. (ed.) *Data and Applications Security XXII*, vol. 5094 of *Lecture Notes in Computer Science*, 283–296 (Springer Berlin Heidelberg, 2008).URL http://dxdoi.org/10.1007/978-3-540-70567-3_22.
12. Sheyner, O., Haines, J., Jha, S., Lippmann, R. & Wing, J. Automated generation and analysis of attack graphs. In *Proceedings. 2002 IEEE Symposium on Security and Privacy*. 273–284 (2002).
13. Wu, S. X. & Banzhaf, W. The use of computational intelligence in intrusion detection systems: A review. *Applied Soft Computing* **10**, 1–35 (2010). URL http://www.sciencedirect.com/science/article/pii/S1568494609000908.
14. Manadhata, P. & Wing, J. An attack surface metric. *Software Engineering, IEEE Transac-tions on* **37**, 371–386 (2011).
15. Debar, H., Dacier, M. & Wespi, A. Towards a taxonomy of intrusion-detection systems. *Computer Networks* **31**, 805–822 (1999). URL http://www.sciencedirect.com/science/article/pii/S1389128698000176.
16. Roy, S. *et al.* A survey of game theory as applied to network security. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* 1–10 (2010).
17. Ingols, K., Chu, M., Lippmann, R., Webster, S. & Boyer, S. Modeling modern network attacks and countermeasures using attack graphs. In *Computer Security Applications Conference, 2009. ACSAC '09. Annual* 117–126 (2009).
18. Verizon. 2012 data breach investigations report.
19. DVLabs, H. 2011 top cyber security risks report.
20. Brennan, J. J. *et al.* Using science and technology investment to improve department of defense cybersecurity metrics. (2009).
21. Alpcan, T. & Baar, T. *Network Security: A Decision and Game-Theoretic Approach* (Cam- bridge University Press, New York, NY, USA, 2010), 1st edn.
22. Mell, P., Scarfone, K. & Romanosky, S. A complete guide to the common vulnerability scoring system version 2.0. In *Published by FIRST-Forum of Incident Response and Security Teams*, 1–23 (2007).
23. Frigault, M., Wang, L., Singhal, A. & Jajodia, S. Measuring network security using dynamic bayesian network. In *Proceedings of the 4th ACM Workshop on Quality of Protection*, QoP '08, 23–30 (ACM, New York, NY, USA, 2008). URL http://doi.acm.org/10.1145/1456362.1456368.
24. Demchenko, Y., Gommans, L., de Laat, C. & Oudenaarde, B. Web services and grid secu- rity vulnerabilities and threats analysis and model. In *Proceedings of the 6th IEEE/ACM international workshop on grid computing,* 262–267 (IEEE Computer Society, 2005).
25. Vieira, M., Antunes, N. & Madeira, H. Using web security scanners to detect vulnerabilities in web services. In *Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on* 566–571 (IEEE, 2009).
26. Yang, S., Holsopple, J. & Sudit, M. Evaluating threat assessment for multi-stage cyber attacks. In *Military Communications Conference, 2006. MILCOM 2006. IEEE*, 1–7 (2006).
27. Amezziane, M., Al-Shaer, E. & Ali, M. On stochastic risk ordering of network services for proactive security management. In *Network Operations and Management Symposium (NOMS), 2012 IEEE*, 994–1000 (2012).
28. Stewart, A. J. Distributed metastasis: a computer network penetration methodology. *Phrack Magazine* **9** (1999).
29. Clarke, G. M. *et al.* Basic statistical analysis in genetic case-control studies. *Nat. Protocols* **6**, 121–133 (2011). URL http://dx.doi.org/10.1038/nprot.2010.182.
30. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, un- certainty and challenges. *Nat Rev Genet* **9**, 356–369 (2008). URL http://dx.doi.org/10.1038/nrg2344.
31. Balding, D. J. A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**, 781–791 (2006). URL http://dx.doi.org/10.1038/nrg1916.

Acknowledgments

The authors would like to thank Dr. Hasan Cam, Dr. Engin Kirda and David Blank-Edelman for valuable discussions and fruitful advice. Special thanks to Rajiv Shridhar, Ray Lisiecki and Joseph De Nicolo for their essential support with the data collection. This research was supported by the Army Research Lab (ARL) through Grant W911NF-11-2-0086, under the project "Metrics for monitoring robustness and controllability".

Author contributions

S.G. and A.-L.B. conducted the principal research and wrote the manuscript. S.G. analyzed the data and confected the figures. A.K. contributed seminal ideas and advised on the research and the manuscript. All authors have reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Gil, S., Kott, A. & Barabási, A.-L. A genetic epidemiology approach to cyber-security. *Sci. Rep.* **4**, 5659; DOI:10.1038/srep05659 (2014).